Data organisation practices

Introduction to Data Management Practices course

NBIS DM Team

data@nbis.se

https://nbisweden.github.io/module-organising-data-dm-practices/

CC BY

Objectives

1. Overview of data storage and processing at different stages of a research project
2. Strategies for organising and naming files in folders, and file shares
3. Working with tabular data in spreadsheets
4. Data entry with data validation in spreadsheets
5. Exporting and importing text-based file formats for tabular data

Credit: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

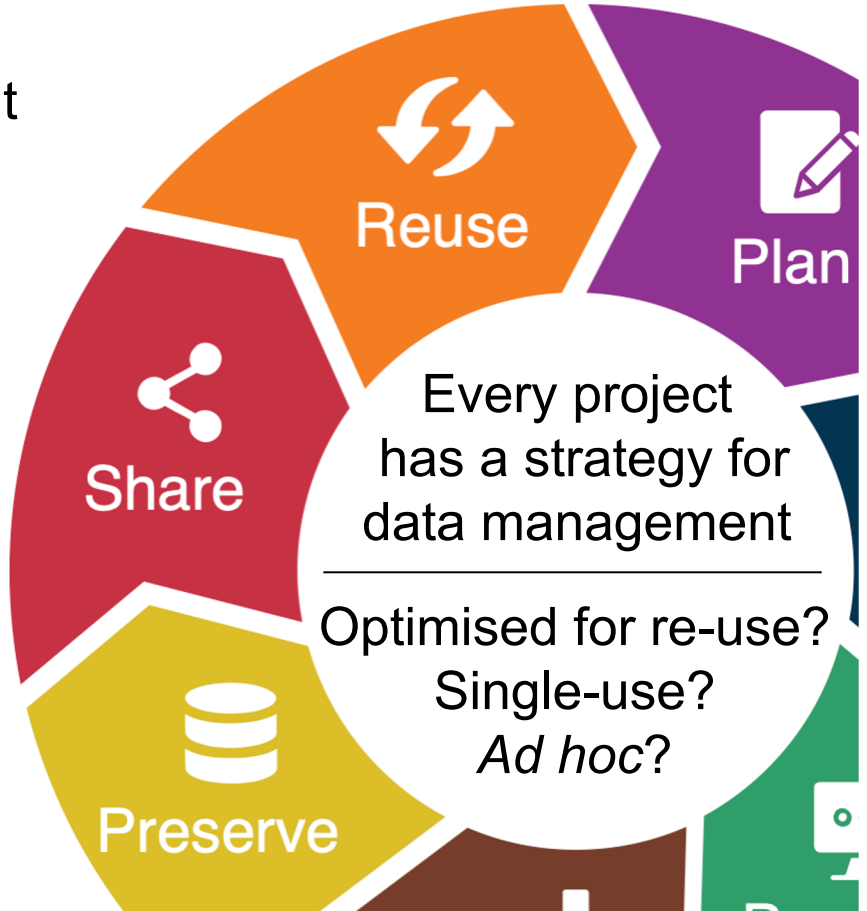
Managing Research Data

- Research **data is a core component** of any research project or publication.
- Good data management practices are **important in all phases** of research
  - Ethics and legislation
  - Information security
  - Research documentation
  - Project organisation
- Research data needs to be secured **beyond the project**'s time frame

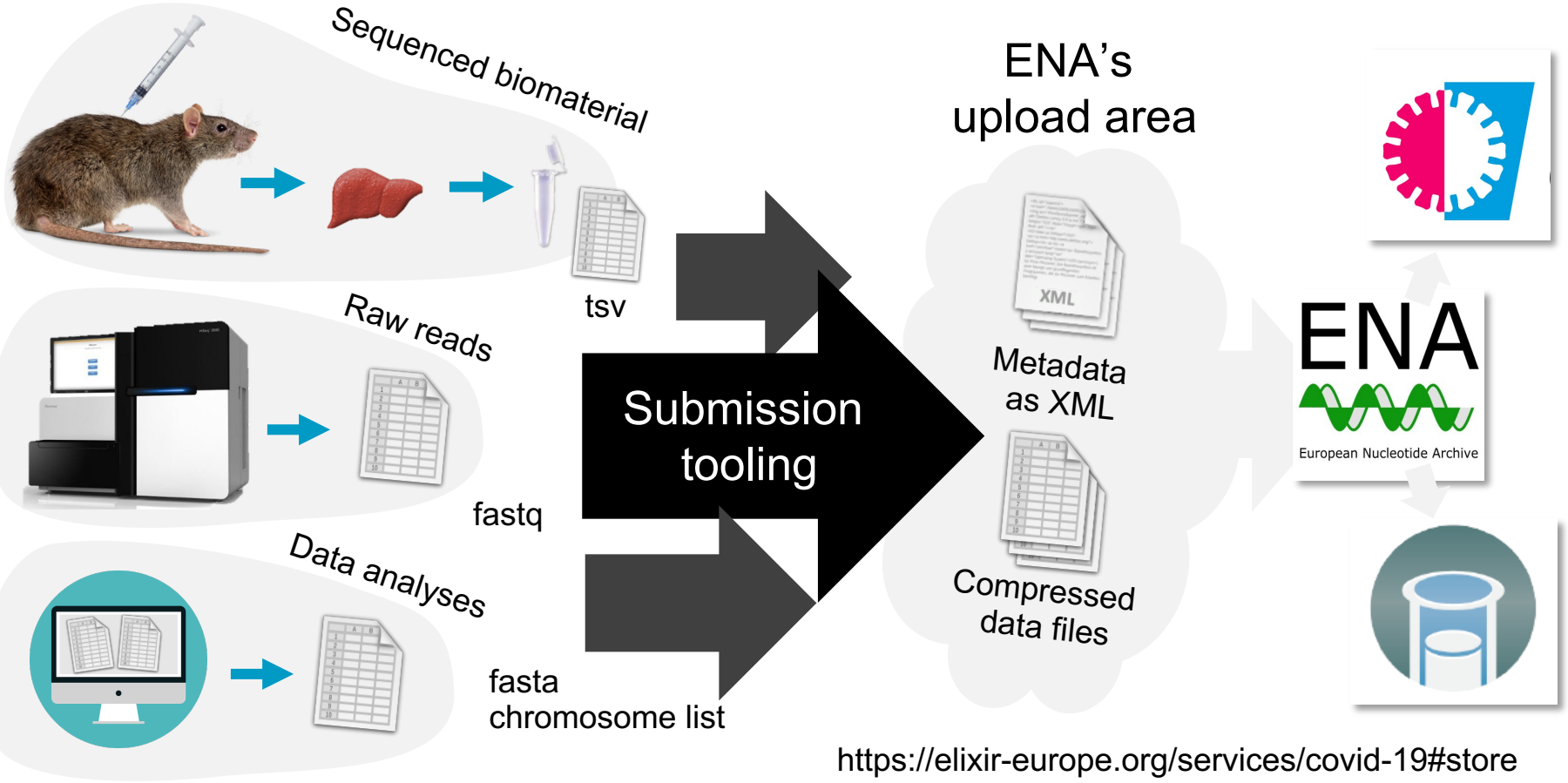
Credit: Illustration from Digitalbevaring.dk / Jørgen Stamp (CC BY 2.5 Denmark license).

Planning for Data Management

- Aim to make your project more efficient by implementing **good practices for handling research data**
- Establish procedures to **organise**, **document**, **preserve** and **share** your research data throughout its life cycle
- Adopt best-practice guidelines that encourage **Reproducible Research**, **Open Science** & **FAIR data principles**



Files and storage



https://elixir-europe.org/services/covid-19#store

Storage & back-up

❑ How to store during the project?

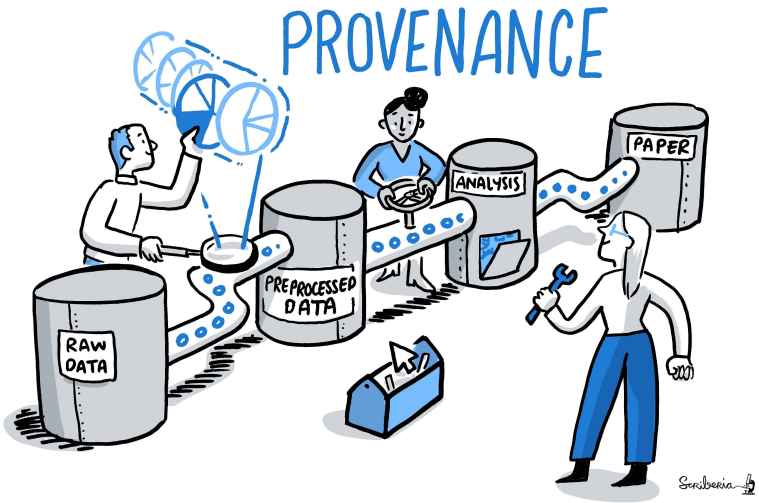
- **Storage/processing locations**
  For data collection, analysis, reporting, code, transfers etc.
- **Back-up and data recovery**
  Strategies to mitigate risks of data-loss and data corruption?
  (Beware of laptops and external storage)
- **Technical requirements**
  Software and systems required to access / process the data?

❑ How to protect data?

- **Information classification**
  Suitable storage based on the characteristics of the data?
- **Access control**
  Who will have access to what data and how will it be enforced?
- **Data protection procedures**
  Other strategies to mitigate risks of unwanted data disclosure or sabotage.

Data storage and processing

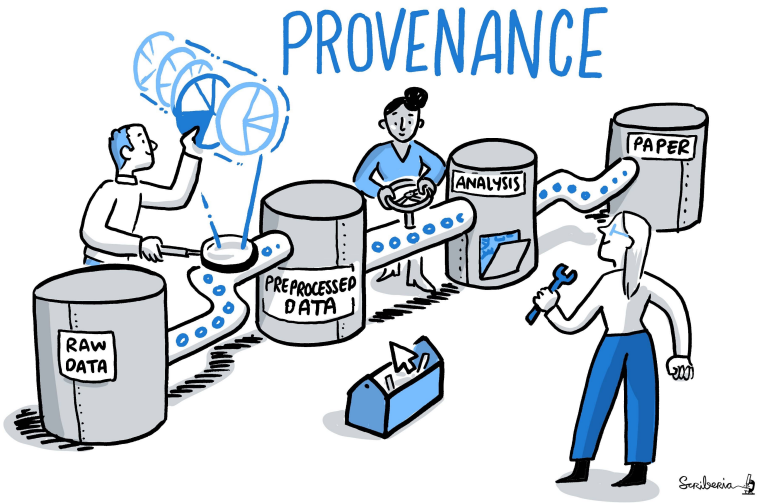
- Data has a life cycle
  - Raw (experiment) data – produce, collect, license, get access, …
  - Processed – generate, clean, aggregate, label, transform, analyse, …
  - Long-term storage – document, select, convert, package, submit, …
  - Published – FAIRify, promote reuse, …
- Maintain data integrity and authenticity
- Plan a storage strategy
- Plan a backup and disaster recovery strategy



Credit: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

Discussion 1

What do you do to **maintain data integrity and authenticity in your projects** and what else could you be doing?



Credit: This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence.

Managing Research Data

The data lifecycle

Data integrity and authenticity

- **Keep original (raw) versions** of data files, or keep documentation that allows the reconstruction of original files
- **Track the location of files** if they are stored in a variety of locations
- Establish **terms and conditions of data use** within the project team and beyond
- Keep a ‘master file’ of the data and take measures to preserve its authenticity
- Decide **how many and which versions to keep** for how long
- **Document changes** that were made in any version
- **Record relationships between items** where needed, for example between code and the data file it is run against

(UK Data Service, 2017a; Krejčí, 2014)

Discussion 2

What **storage categories and factors do you consider** when selecting which category to use or not to use?

- Portable devices
- Cloud storage
- Local storage
- Networked drives

Portable devices

- **Temporary, short-term storage for non-sensitive data**, e.g. in the field or to transport data and files when online transmission is not possible.
- In combination with **encryption and strong password protection**, especially if working with sensitive information.

✓ Conduct **regular checks** to ensure your device is working and that files are accessible.

❖ **Not for long-term storage or master copies** of your data

Cloud storage

- Granting **shared, remote and easy access to data and other files** to all involved in the project
- **Read the terms of service**. Especially focus on rights to use content given to the service provider
- **Opt for European, national, or institutional** cloud services which store data in Europe if possible

- **Not your only storage and backup solution**
- **Not for unencrypted (sensitive) personal data**

Local storage

- When working on different (local) workstations, e.g. laptop at home and the desktop in the office:
  - **always make sure that you are working on the most current version**, for example with the help of versioning software or guidelines
  - make sure that the most **current version is always backed up somewhere else**

✓ Suitable as a primary storage for projects involving only very few people

❖ Avoid if data will be moved back and forth between personal computers frequently

Networked drives

- **Use in projects involving many people** who need access to data and files
- Use a suitable security strategy to **protect data and files against unauthorised access**
- **Agree on rules for versioning files and data** to ensure that everyone can locate and access what they need
- **Long-term store complete data** that has been analysed, which can be cost efficient and offer increased security
- Restrict access where possible using rights and permissions, e.g. **write protect a master copy** and only grant access to specific files/folders when necessary

Discussion 3

What are examples of potential causes for data loss in a research project?

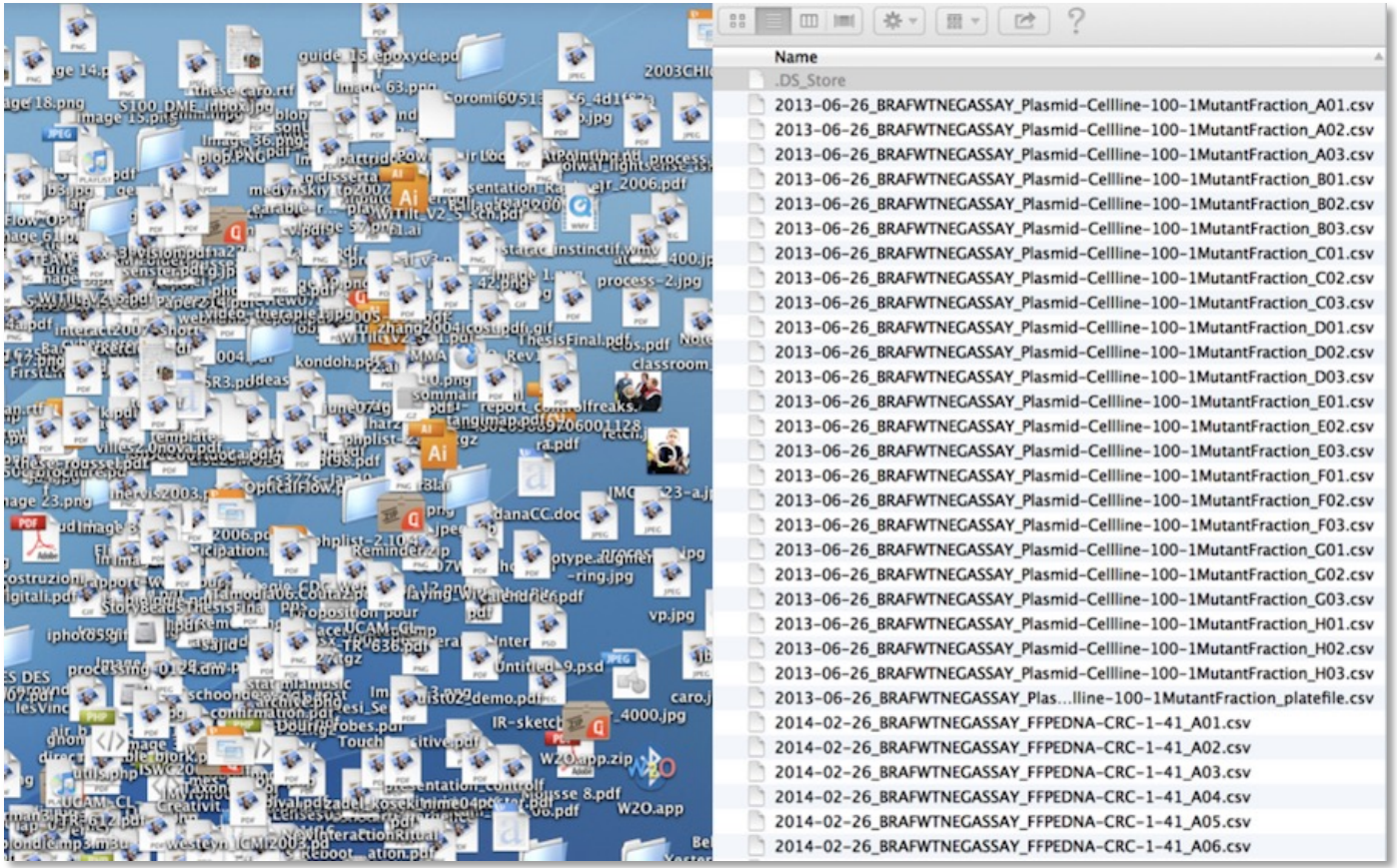
Backup and disaster recovery

A minimalist strategy

- ❑ There are at least
  - ➢ Three copies of the data, of which…
    - ✓ Two are kept on different types of storage media
    - ✓ Two are at different locations
    - ✓ One is located off-site
- ❑ All copies are checked regularly to make sure that they work
- ❑ The process is known and applied in the project (automated)

… also determine what you want to back up, and find out whether your institution already has a backup strategy.

Organising files and folders



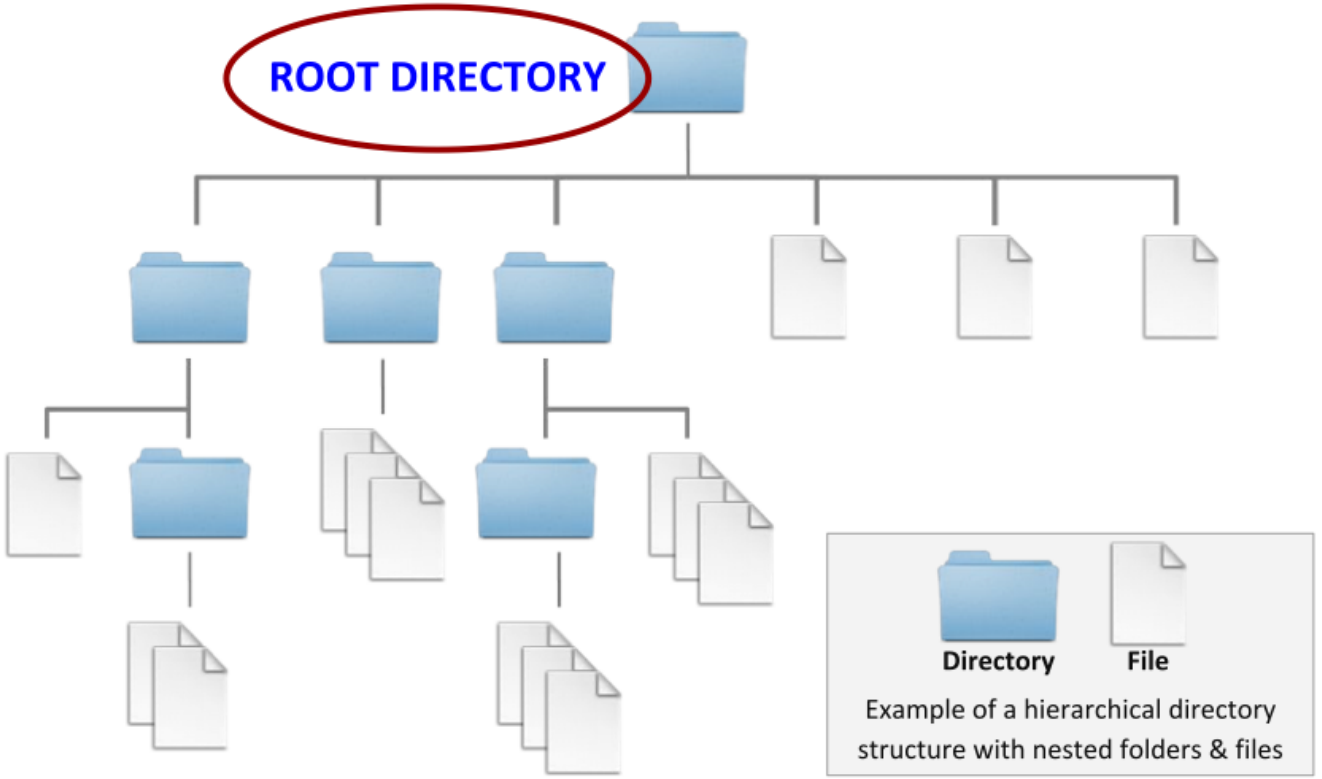
Practices

Your primary collaborator is yourself from 6 months ago, and that person refuses at answer your e-mails!

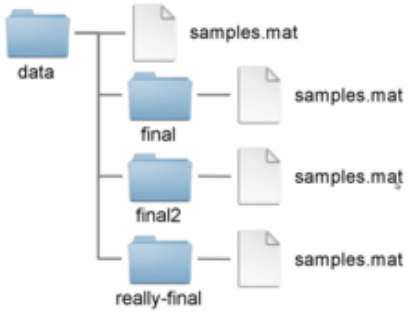
Good practices

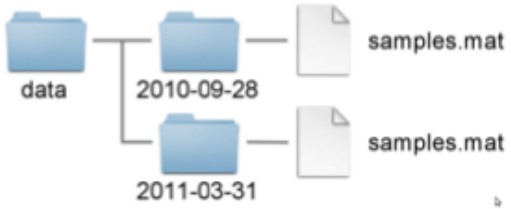
- Organise files hierarchically
- Use folders to divide files into categories
- Choose a file naming strategy
- Create documentation files

Organise files hierarchically

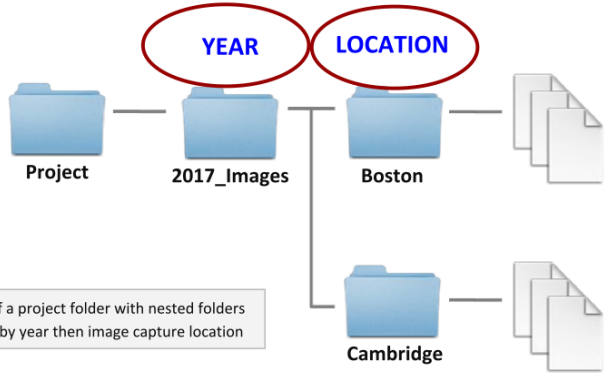


Divide files into categories



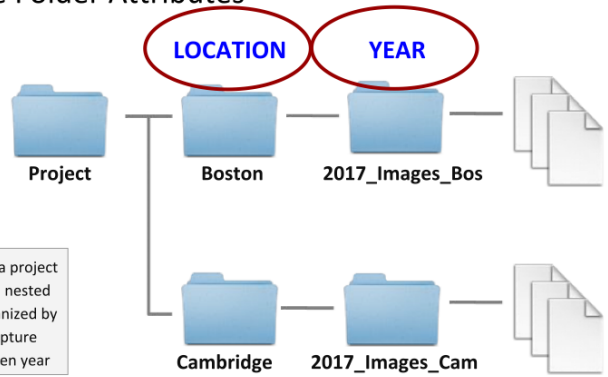


Create Folder Attributes



Example of a project folder with nested folders organized by year then image capture location

Create Folder Attributes

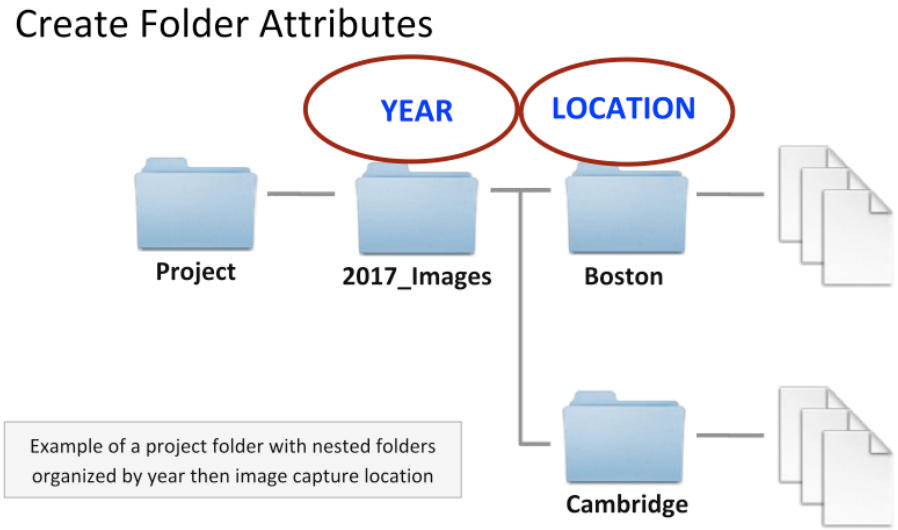


Example of a project folder with nested folders organized by image capture location then year

Discussion 4

What characteristics do you use to create folders and subfolders in your projects?

E.g., data, collection year, ...



Helpful characteristics

- Year or other date
- Type of data, document or file
- Project stages
- Analysis version or revision
- Experiments
- Instruments
- Time periods
- Geographic location
- Storage requirements
- Team member, institution or project site

Choose a file naming strategy

- A file name is a principal identifier for the file
- Consistent in time and among different people
- Practically useful when accessing files, such as sorting and filtering

Chronologically
(ISO 8601 date standard)

20171028_001.tiff
20171028_002.tiff
20171028_003.tiff
20171029_001.tiff
20171029_002.tiff

Classification or code
(standardized)

USNM_379221_01.tiff
USNM_379221_02.tiff
USNM_379221_03.tiff
USNM_379222_01.tiff
USNM_379222_02.tiff

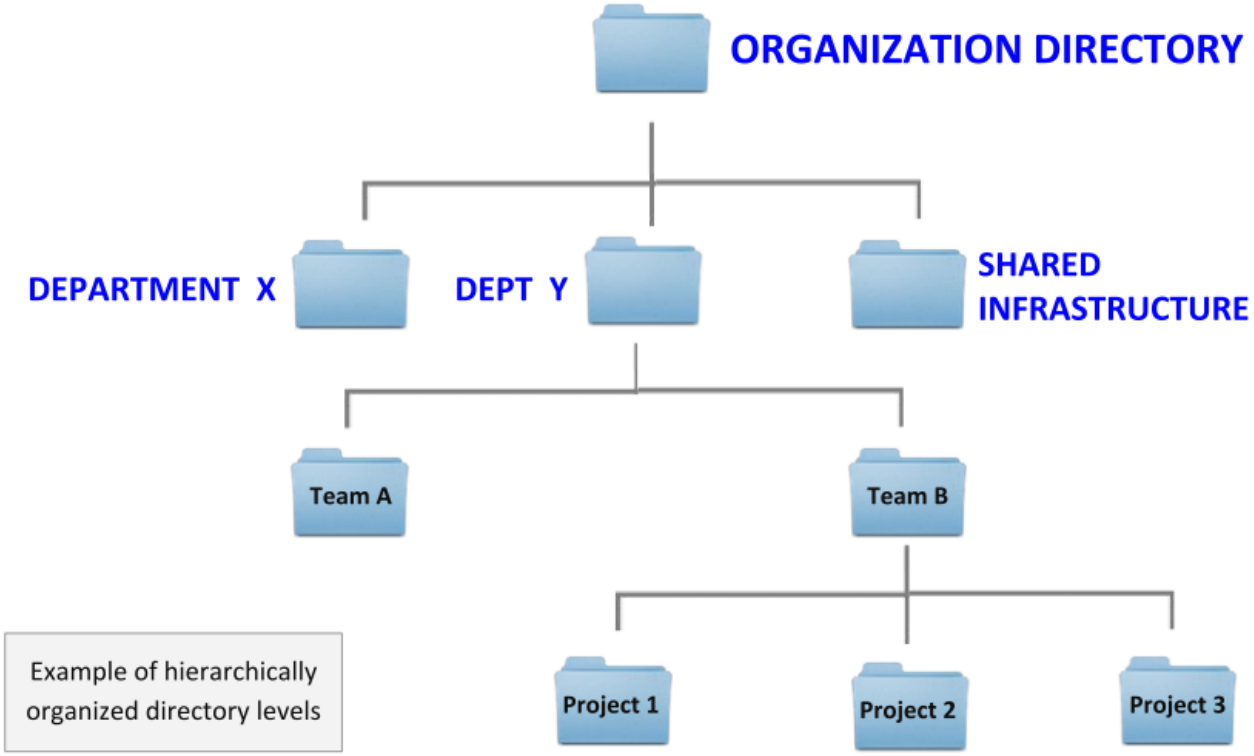
Alphabetically
(depending on type of files)

bos_20171028_001.tiff
bos_20171028_002.tiff
bos_20171029_001.tiff
cam_20170922_001.tiff
cam_20170922_002.tiff

Discussion 5

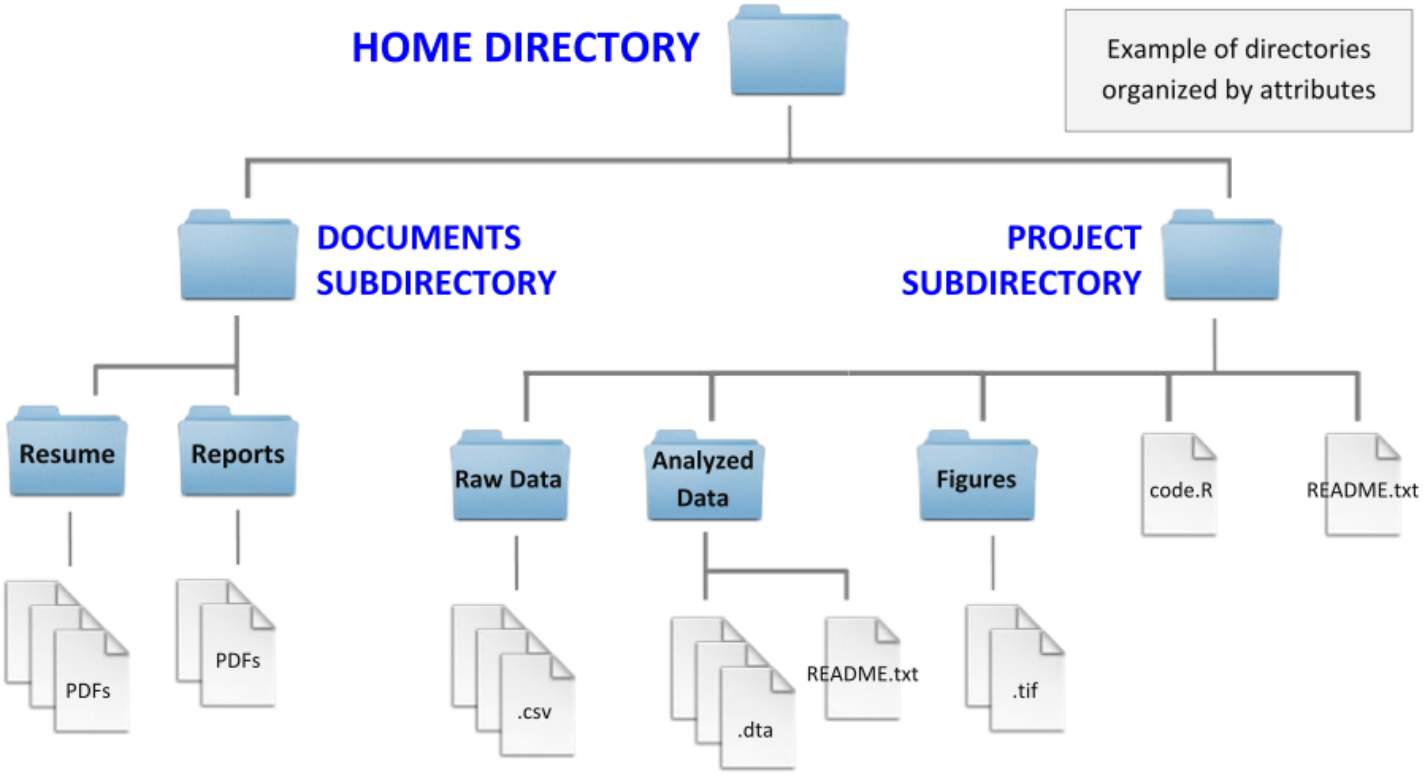
What are examples of potential benefits of agreeing on a File Naming Convention for a project?

Hierarchy?

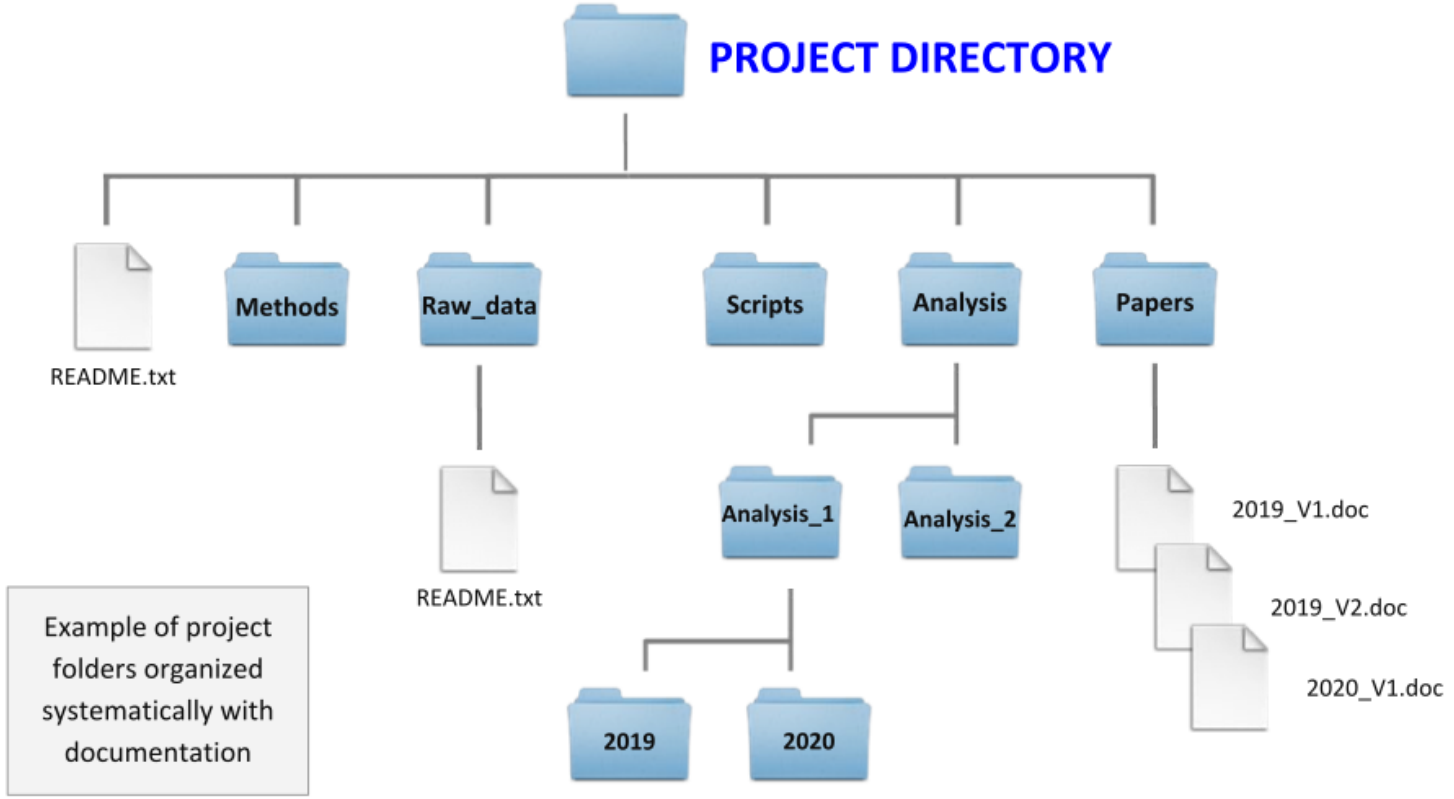


Example of hierarchically organized directory levels

Hierarchy?

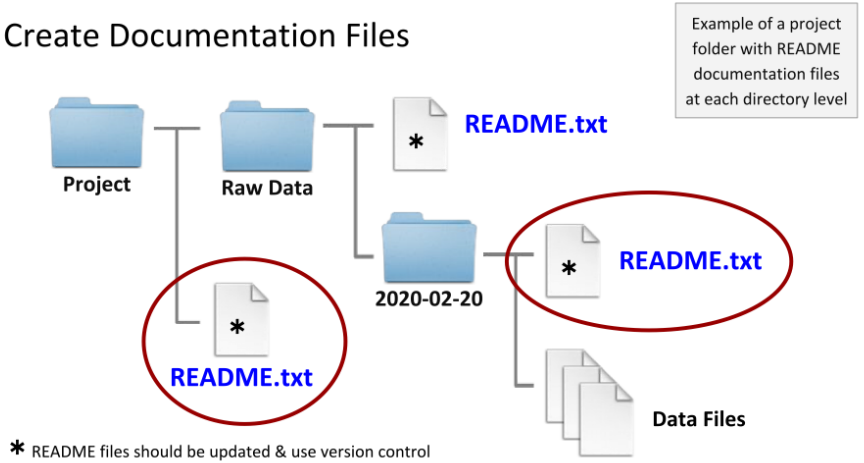

HOME DIRECTORY
Example of directories organized by attributes
DOCUMENTS SUBDIRECTORY
PROJECT SUBDIRECTORY
Resume
Reports
Raw Data
Analyzed Data
Figures
code.R
README.txt
PDFs
PDFs
.csv
.dta
README.txt
.tif

Hierarchy?


PROJECT DIRECTORY
README.txt
Methods
Raw_data
Scripts
Analysis
Papers
README.txt
Analysis_1
Analysis_2
2019_V1.doc
2019_V2.doc
2020_V1.doc
2019
2020
Example of project folders organized systematically with documentation

Create documentation files

- Key to making the data publishable, discoverable, citable and reusable
- Project-level and data-level documentation
- Use plain text files and save them in your file hierarchy – README-files

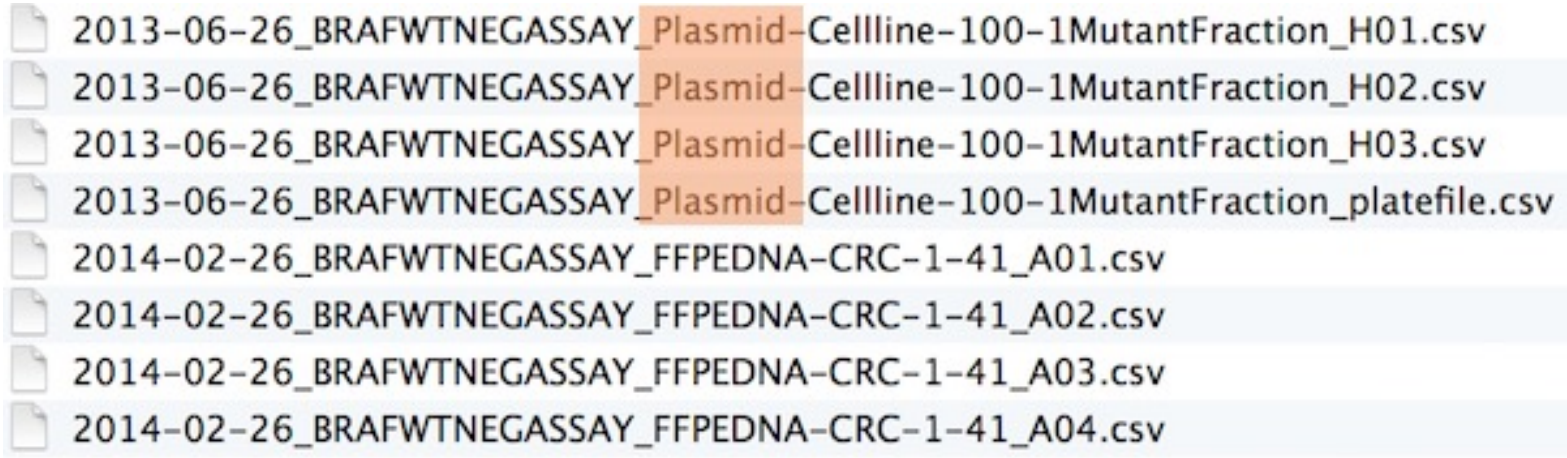


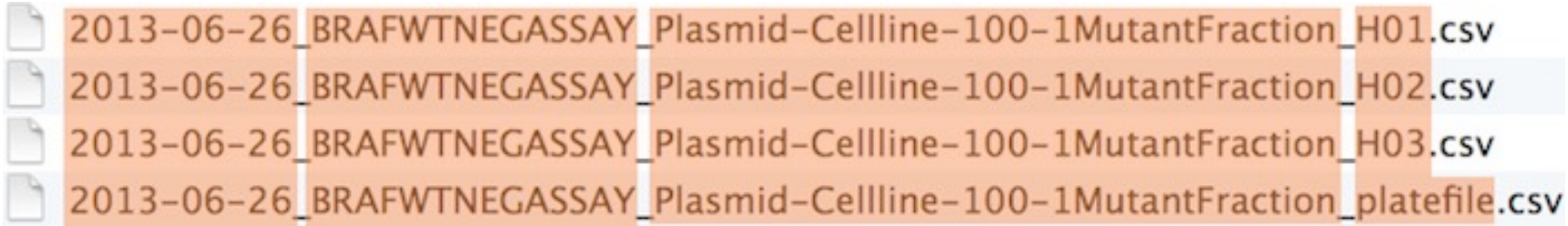
Human readable

```
01_marshal-data.md
01_marshal-data.r
02_pre-dea-filtering.md
02_pre-dea-filtering.r
03_dea-with-limma-voom.md
03_dea-with-limma-voom.r
04_explore-dea-results.md
04_explore-dea-results.r
90_limma-model-term-name-fiasco.md
90_limma-model-term-name-fiasco.r
Makefile
figure
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
tmp.txt
```

```
01.md
01.r
02.md
02.r
03.md
03.r
04.md
04.r
90.md
90.r
Makefile
figure
helper01.r
helper02.r
helper03.r
helper04.r
tmp.txt
```

Machine readable





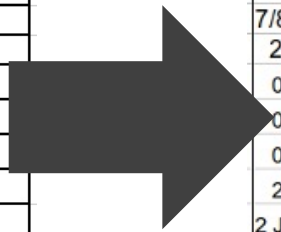
Naming files and folders

- ✓ Optimise for your needs
  - Locating files and preventing mistakes
  - Keeping track of milestones / versions
  - Exploration, e.g., with readme, license, data dictionary files, etc.
- ✓ Be consistent
  - Agree on conventions early but be prepared to revisit and revise
  - Encode useful metadata in names of files and folders
  - Choose names that facilitate sorting and filtering
- ✓ Use folders to avoid clutter
  - Organise files hierarchically by project, task, and/or type:
    e.g., data_raw, data_output, scripts, docs

Discussion 6

What structural issues do you often find in spreadsheets and how do you address them?

RDM training			
Date	Length (hours)	PGR|PDRA|other	Delivered by
4 Feb	1.5		GQ
7/8 Feb			GQ
20 Feb			GQ & DF
03/03/17	2	15|03|00	DF
04/03/17	2	30|0|0	DF
08/04/17	2	30|0|1	DF
26/05/17	2	27|0|0	DF
2 June?	2	24|02|00	DF
3 June?	1.5	12|07|04	DF



RDM training					
Date	Length (hours)	PGR	PDRA	other	Delivered by
4 Feb	1.5				GQ
7/8 Feb					GQ
20 Feb					GQ & DF
03/03/17	2	15	3	0	DF
04/03/17	2	30	0	0	DF
08/04/17	2	30	0	1	DF
26/05/17	2	27	0	0	DF
2 June?	2	24	2	0	DF
3 June?	1.5	12	7	4	DF

Tabular data in spreadsheets

- ✓ Raw means raw!
- ✓ Tidy data tables

  One cell–one value
  One column–one variable
  One row–one observation

- ✓ Beware of Excel “features”

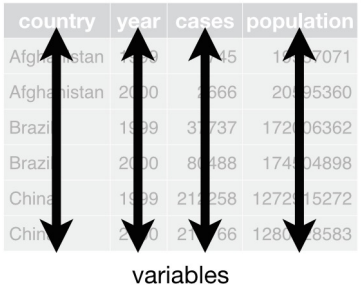
  Misguided ”auto-corrections” of dates, casing, numbers etc.

  Misaligned formulas

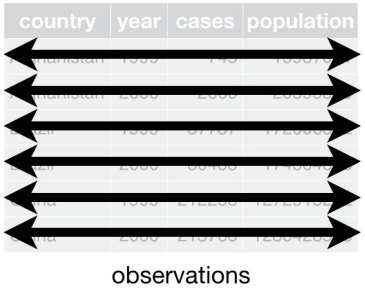
  Limited numerical precision

  Limited number of rows/columns

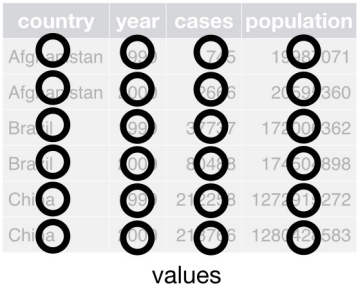
	A	B	C	D	E	F	G	H	I	J	K
1	**data**							analysis			
2	id	biomarker1	biomarker2	biomarker3	biomarker4			variation	ave	problem	
3	81	0.08502	0.07002	0.07735	0.07746			0.008	0.0775		
4	82	0.0658	0.06859	0.06958	0.06799			0.002	0.068	no	
5	83	0.07757	0.07497	0.0801	0.07755			0.003	0.0775		
6	84	0.07185	0.06957	0.07474	0.07205			0.003	0.0721	yes	
7	85	0.06959	0.07361	0.07113	0.07145			0.002	0.0714	maybe	
8	86	0.09291	0.10439	0.09425	0.09718			0.006	0.0972		
9	87	0.07878	0.08143	0.07203	0.07742			0.005	0.0774		
10	88	0.07907	0.077	0.08227	0.07944			0.003	0.0794		
11	89	0.07299	0.07616	0.08131	0.07682			0.004	0.0768		
12	90	0.07487	0.0664	0.0671	0.06946			0.005	0.0695		
13											
14	mean	0.076845	0.076214	0.076986	0.076682						
15								biomarker QC			
16	notes							b1	b2	b3	b4
17	* patient id86 may need removing due to missing notes							0.46336967	0.875281336	0.918250702	0.14953926



variables



observations



values

Data validation in Excel

- Data value types
- Value constraints
- Data dictionary (metadata module)

	A
1	species
2	
3	Dipodomys merriami
4	Dipodomys ordii
5	Dipodomys spectabilis
6	

	A	B	C	D	E
1	**Field name**	**Description**	**Data type**	**Units**	**Notes**
2	Name	Full patient name including title	Text	na	From patient notes
3	Address	Full patient address	Text	na	From patient notes
4	Age	Age in years	Numeric (interger)	na	From patient notes
5	Symptoms	Summary of symptoms from last visit	Text	na	Around half of patients have missing notes
6	CRP	C-reactive protein measurement	Numeric (decimal)	mg/ml	Measured on instrument LAB-C-002. Normal range <10

Discussion 7

Can you think of information that may be lost when converting data between file formats?